Final Project Report
Network Aggregation of YouTube Data

Tristan Bailey,
Ean Jacob Gayban,
Gabriel Mortensen
University of Nevada, Reno. CS 431/631
Introduction to Big Data
R11 - Team 1

***Abstract*— This document presents the final project report for the Introduction to Big Data project undertaken by Team 1. This project aimed at exploring the potential of social media as a valuable source of information for promoting social change, particularly in the context of health and well-being. The project leveraged data analysis techniques to investigate societal trends and utilized video statistics from Simon Fraser University's YouTube dataset crawled in 2007. The project was executed in two phases. The initial phase was centered on developing a command-line interface (CLI) integrated with PySpark for data analysis and a Bin Generation function to enable users to input their data and perform analyses. For the final phase, the team utilized SQL to obtain more detailed data, and the R programming language was used for visualization, which resulted in several visualization segments, such as histograms, scatterplots, and pie charts. The team's findings revealed that the Music category accounted for 37% of YouTube videos in 2007 while DIY videos had a small proportion of significantly longer videos. Additionally, UNA (unassigned) videos received over seven times more views than other categories, and Comedy videos of around 150 seconds (give or take 15 seconds) emerged as the most trendy. In conclusion, this project demonstrates the potential of social media data as a valuable resource for driving social change. By utilizing data analysis and visualization techniques, the team was able to uncover several societal trends in 2007, providing valuable insights into the usage patterns and preferences of YouTube users.**

Keywords—component: SQL, PySpark, R, Youtube, Machine Learning, Data Visualization.

I. INTRODUCTION

Social media platforms have revolutionized the way people consume and share information, with YouTube being a sector leader. One article from the Journal of Medical Internet Research notes that "social media remains relatively untapped as a source of information to catalyze policy action and advance social change" [1]. As the amount of data on social media platforms continues to grow, businesses and researchers have increasingly turned to data analysis techniques to explore societal trends and patterns as well as to grow their businesses. In this project, our team aimed to leverage the potential of YouTube as a source of information for promoting social change, specifically in the context of health and well-being.

The project focused on analyzing video statistics from Simon Fraser University's YouTube dataset crawled in 2007 [2] using PySpark, SQL, and R. The project was executed in two phases: developing a command-line interface (CLI) integrated with PySpark for statistical generation as well as utilizing SQL to obtain more detailed data and the second phase which used information obtained from the first phase and R to perform visualization. The team's findings revealed several societal trends in 2007, including the most popular categories being Comedy, Music, and Entertainment. UNA (unassigned) videos receive over seven times more views than other categories. Additionally, the team discovered that DIY videos had a small proportion of significantly longer videos. Finally, Comedy videos of around 150 seconds emerged as the most trendy.

In this paper, we provide an overview of our project, including the data analysis techniques used, the evaluation process through data visualization, and the findings. We demonstrate the potential of social media data as a valuable resource for driving social change and show how data analysis and visualization techniques can be utilized to uncover societal trends and preferences of YouTube users. Our project provides insights into the usage patterns of YouTube users in 2007, providing valuable information for researchers and policymakers interested in understanding the impact of social media on society. For instance, by measuring trends, potential trends could be theorized and potentially thwarted if deemed unsafe (e.g., tide pod challenge convinced thousands to ingest toxic chemicals [3]). The paper concludes by highlighting the significance of our findings and the potential for future research in this area.

II. BACKGROUND/MOTIVATION

A. Apache Spark

Apache Spark is a data processing framework capable of leveraging compute clusters to process very large amounts of data. While we are not using a compute cluster to handle our data, we are taking advantage of its APIs to more easily read, write, and process our dataset.

B. PySpark

PySpark provides access to Spark's functionality by exposing it through a Python library. This makes it easier to work with since the team is familiar with Python. This also gives us the opportunity to use other well-known data processing libraries like NetworkX and Matplotlib, should we choose to use them.

C. Python

Python is an object-oriented programming language that is widely used in the field of data science and analytics. As such, it has numerous open source libraries that allow for the seamless usage of several Big Data systems: such as PySpark and PyTorch. Where these libraries are typically written in a more optimized language, like C, and as such have good compute performance.

D. R

R is a popular functional programming language used in the fields of Statistics and Data Science. It has powerful inbuilt libraries that enable its users to easily generate statistical information from a dataset and to generate elegant graphics from datasets. For this reason this language has been chosen as an additional tool for the project, over using the graphing libraries in Python.

E. YouTube Dataset

This is the dataset that will be analyzed for this project. Researchers from Simon Fraser University in Canada performed several crawls of the YouTube website between 2007 and 2008. For each video they came across, they recorded several pieces of information about it, such as the uploader, views, length, rating, and the related videos that showed up in the sidebar. Because each video is related to several other videos, we are able to reconstruct a graph of the YouTube video network at the time of the crawl. The statistics shown in this paper come from the first crawl on February 22, 2007. This crawl contains data for exactly 749,361 unique videos. The crawl was implemented as a breadth-first search, initially seeded with the videos from the "Recently Featured", "Most Viewed", "Top Rated" and "Most Discussed", for "Today", "This Week", "This Month" and "All Time" sections. When a video is accessed, links to several other videos may be present in the sidebar. The IDs for each video are recorded, and will be processed in the next stage of the crawl.

When processing a video, both information from the video's webpage and from the YouTube API are combined. Table I shows the information that was recorded for each video.

Data collected for each video	
video ID	a unique 11-digit ID string
uploader	the uploader's username
age	the age of the video
category	category chosen by the uploader
length	the length of the video, in seconds
views	the number of views
rate	the video rate, stored in a float
ratings	the video rating stored as an integer
comments	the number of comments
related IDs	up to 20 related video IDs

Table I: Table showing each piece of information obtained for a video, alongside an explanation of what it represents.

F. Degree Distribution

The degree of a vertex in a graph represents the number of connections or edges k it has to other vertices. The degree distribution $P(k)$ of a graph is defined as the proportion of nodes with degree k. If there are n_k nodes with degree k, then $P(k) = n_k/n$. For directed graphs, there are two measures of degree: in-degree, which measures the number of inbound edges (other vertices point to this one), and out-degree, which measures the number of outbound links (this vertex points to others). The degree distribution can easily be plotted using a histogram, with degree k on the x-axis and $P(k)$ on the y-axis. Alternatively, we can also plot the number of occurrences on the y-axis. Using this data, we can easily determine statistics such as the average, minimum, and maximum degree k. For directed graphs, we can generate these statistics for both the in- and out-degree.

III. Approach

A. Degree Distribution

By considering each video ID as a vertex, we are able to treat the list of related IDs as the "adjacency list" for that video. This way, the dataset implicitly defines a graph that can be used to determine the degree. Because each (video ID, related ID) pair represents a directed edge, we can calculate both the in-degree and out-degree for all videos.

Calculating the out-degree is particularly simple. Because we have an adjacency list, the out-degree represents the length of that list. Normally, each data file provided by Simon Fraser University will contain only 1 record of each video. If we wish to use multiple data files in our program, we would have to determine if there exists two or more records for the same video, then make sure to merge the two adjacency lists in each record. In this case, the out-degree would be the length of this merged adjacency list.

To calculate the in-degree, a few processing steps are needed since PySpark does not provide many options to operate on lists. To get around this issue, we converted each (vertex, list) pairs into multiple (src vertex, dst vertex) pairs. The in-degree is defined as the number of vertices that point to a particular vertex. In other words, the in-degree of a vertex is the number of times it is used as a destination vertex. Because we transformed our dataset into a set of (src, dst) pairs, we can identify all the unique destination vertices, then count how many times they occur in the *dst* column.

To find the overall degree of a video, we can add the in-degree to the out-degree. Once we've calculated these values, we can then count how many times each degree value occurs in the dataset, then display a distribution. For our application, we opted to export the raw degree data itself, to be visualized in R.

B. Categorized Statistics

The command-line interface can calculate a distribution for all of the given fields, not just the degree. For example, it can determine which categories are the most common, and what the most common video length is. Generating statistics for the categorical fields (i.e. *uploader* and *category*) is straightforward, just count the number of times each value shows up. There are a finite amount of uploaders and categories throughout the dataset, so the distribution is easy to work with.

When creating a distribution for the numerical fields (*age*, *length, views, rate, ratings,* and *comments*), we opted to require "buckets" before generating the distribution. This allows the user to define the level of detail they want for their application. For example, one might use logarithmic buckets like [1,000-10,000), [10,000-100,000), [100,000-1,000,000), and so on to group the number of views a video has.. If they wanted a more detailed look at the video distribution, they might shrink the size of the buckets.

The command-line interface allows for a filter to be applied to the dataset to allow for more granular statistics. For example, the user might only care about videos in a particular category, or they might want to focus on the videos that have more than 1,000,000 views. This filter is applied as a SQL WHERE clause. Only 1 filter may be applied on the dataset.

The comment-line interface is able to generate statistics for multiple columns at the same time. For example, if the user wishes to group the data by category, then by views, then the program will generate a view distribution for each category. This is useful to identify any relationships between each of the tracked variables.

C. Exporting Data

If the user needs to generate more complex statistics from the dataset, then the command-line interface provides an export feature to return the entire dataset as a CSV file. The raw data provided by Simon Fraser University is poorly formatted, so this export feature also doubles as a cleaning tool for the dataset. The user is able to select which columns they want to export if they only need a subset of the data. This exported data can then be used by other programs. The figures shown later in the paper make use of the exported CSV data from the command-line interface.

PySpark allows for SQL statements to be executed on the dataset. The command-line interface exposes this functionality to the user, allowing them to write a SELECT statement on the entire dataset. This allows the user to access SQL operations like GROUP BY and most of the basic aggregation functions. This allows for some preprocessing of the data before export, or just to test the user's queries against the dataset.

D. Data Visualization

If the user needs to generate more complex statistics from the dataset, then the command-line interface provides an export feature to return the entire dataset as a CSV file. The raw data provided by Simon Fraser University is poorly formatted, so this export feature also doubles as a cleaning tool for the dataset. The user is able to select which columns they want to export if they only need a subset of the data. This exported data can then be used by other programs. The figures shown later in the paper make use of the exported CSV data from the command-line interface.

PySpark allows for SQL statements to be executed on the dataset. The command-line interface exposes this functionality to the user, allowing them to write a SELECT statement on the entire dataset. This allows the user to access SQL operations like GROUP BY and most of the basic aggregation functions. This allows for some preprocessing of the data before export, or just to test the user's queries against the dataset.

V. Evaluation and Data Visualization

For this portion of the project the team had to take a step back and consider what the actual study was about. Our team wanted to best track the patterns of social trends via the analysis of media. After much deliberation it was decided that the team should frame the data visualization portion of this project around various questions. This section will go over answering each of the questions presented by the team and will be accompanied by an analysis of the graphic used to provide said answer.

A. What were the most viewed videos like in 2007?

The team chose this question because by focusing on the majority, the team believes that they can gain insights into overall societal views and behaviors. This is because social media platforms are used by a large portion of the population, and therefore the data generated can provide a representative sample of broader societal trends [X]. Additionally, by analyzing the views of the majority, the team can identify patterns and trends that are more likely to be influenced by societal norms and values, as these perspectives are shared by a larger number of people. To answer this question the team generated two graphs which can be observed in Figure I and Figure II.

From the figures presented the team was able to identify the most popular categories as Music, Entertainment, and Comedy. Music was rated the highest with 27% of the overall view count. Categories that were less than the top 3 percentages were placed in their own category called "Other". Having identified the most popular video the team then generated a bar graph representing the average length time of each category to which it was discovered that the top three categories were all clustered together around the same length, that is between 200 and 250 seconds.

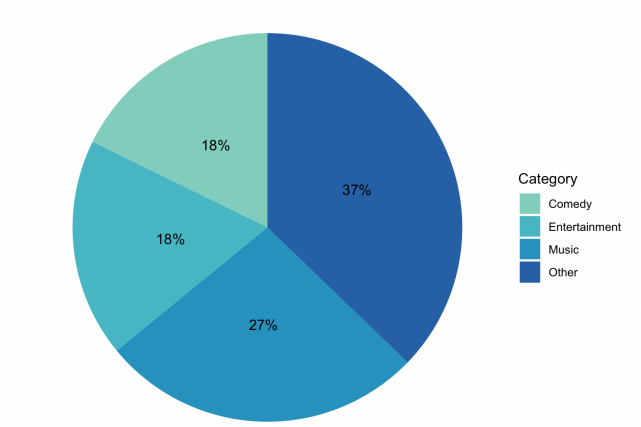

Figure I: Pie chart showing overall view count by category.

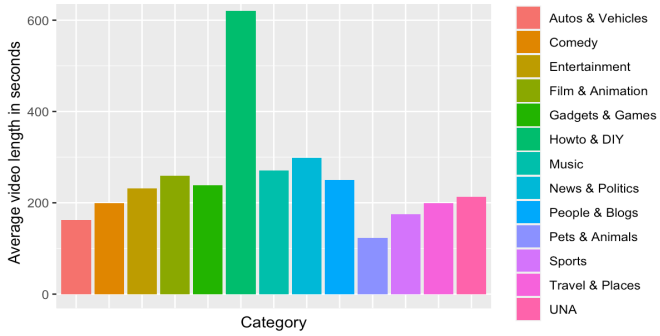

Figure II: Bar graph representing average length of videos by category. The longest videos were Howto & DIY with roughly 600 seconds of average coverage.

B. What do the videos with the highest degree look like?

Since the goal of this project was to examine social trends the team decided to place their efforts on examining Backlinks. Backlinks are the number of incoming links to a video (i.e., a digital link leading to content). Theoretically the more links a video has the more likely it is to get watched and thus make an impact to society. To determine the videos with the highest degree the team first performed a general overview of backlinks and their existence as shown in Figure III.

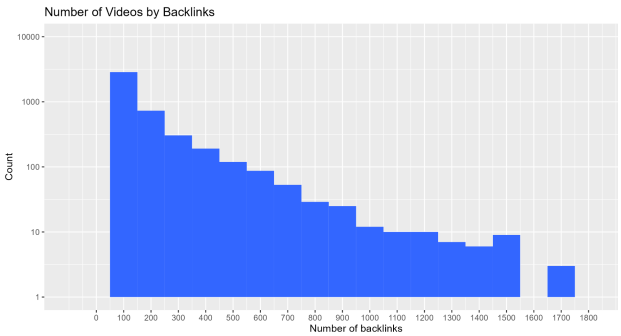

Figure III: Histogram pertaining to how many videos have a particular number of backlinks.

Figure III shows that backlinks decline at an exponential rate which could imply a hierarchical structure of popularity. An exponential decline in backlinks could represent a hierarchical structure of popularity because it suggests that a few videos have a large number of backlinks while the majority have fewer links. This is consistent with the notion of a power-law distribution.

Power law distributions are often observed in complex systems and are believed to indicate the presence of underlying complexity in the generating process [4]. This project is comparable to that of a social network, for example, power law distributions are often observed in the distribution of node degrees, where a few nodes have a large number of connections while the majority have fewer connections. This can be seen as a signature of the complex interplay between individual behavior, network structure, and external factors that shape the evolution of the network [5].

Figure IV is another graphic generated by the team to further understand the correlation of degrees. In this case the team focused on comments rather than views. Studying the number of comments in video extracted can provide valuable insights into social trends because it reflects the level of engagement and interest among viewers. Videos with a high number of comments suggest that the content resonated strongly with the audience, potentially reflecting popular opinions, interests, or concerns. By analyzing the topics and sentiments expressed in the comments, the team could gain a deeper understanding of social phenomena. Moreover, the popularity of videos and the number of comments could be used as a measure of influence since highly commented videos are likely to have a greater impact on their audience and potentially contribute to the formation of social norms and behaviors.

Unexpectedly, the team observed that the degree count within a video was not nearly as effective as the view count when correlated to the frequency of comments. This could mean that despite the potential insights that can be gained from analyzing the number of degrees within a video, views remain the more reliable metric for assessing a video's impact and popularity in 2007. Possibly from typical marketing concepts such as word of mouth sharing of the video, as opposed to video advertisements through degree drawing in viewers.

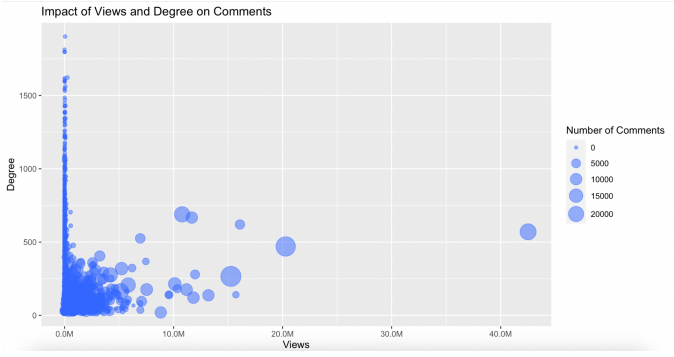

Figure IV: Bubble chart determining the impact degree and views have on the frequency of comments.

C. What was an average video in 2007 like?

The fourth main category of questions the team evaluated for the 2007 Youtube dataset focused on what were the average videos like around that time with respect to categorical statistics. The first pertinent categorical

statistic for youtube videos is the number of views that video received. To perform this evaluation videos were separated into their respective categories. From here video views were rounded to the nearest 100. Then these categories were visualized with respect to other categories when it came to average and median views. These are depicted in Figure VI and Figure V respectively.

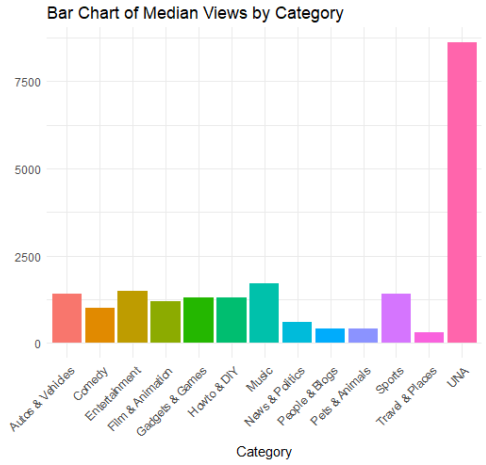

Figure V: Bar Chart of Median Views for Each Category.

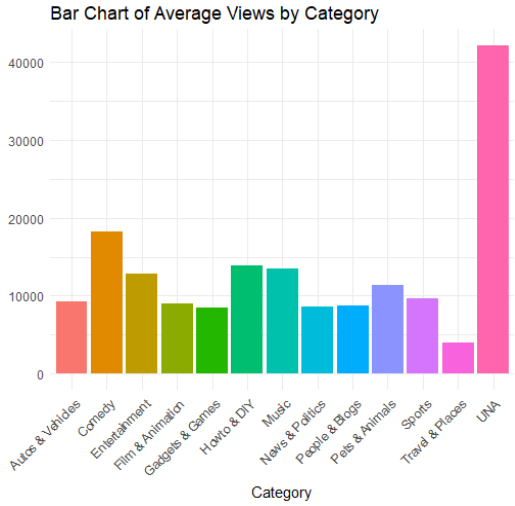

Figure VI: Bar Chart of Average Views for Each Category.

From these two figures the first observation is that UNA videos both have a higher median and average number of views indicating that this category overall typically receives higher viewership than the other categories. The second observation is that as for all categories they typically have an average viewership significantly higher than median, indicating that there are outlier videos in each category with several factors more views than the typical video. Indicating that every category has viral videos.

The next categorical statistic used to evaluate the average video in 2007 was the video duration. To perform this evaluation the team used the same method as for the views statistics but instead rounded video durations to the 10's of seconds. The average and median views are depicted in Figure VIII and Figure VII respectively.

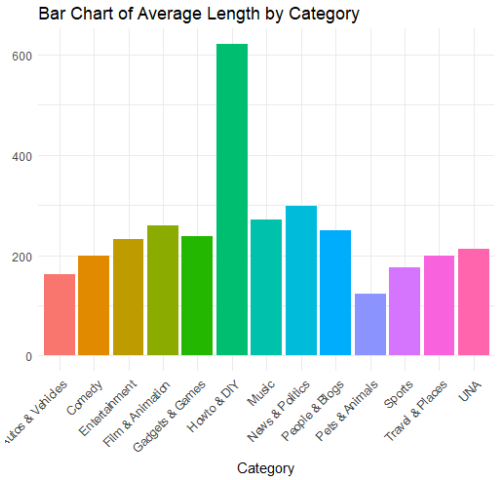

Figure VII: Bar Chart of Average Video Duration for Each Category.

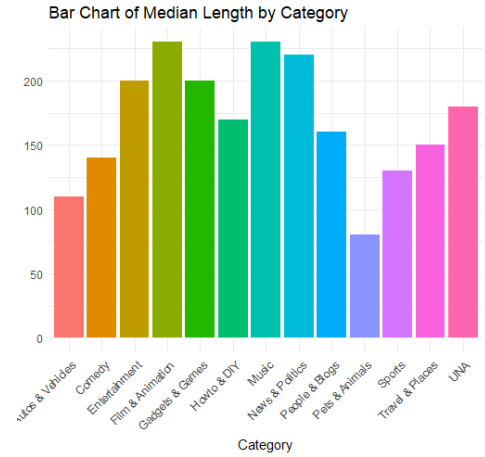

Figure VIII: Bar Chart of Median Video Duration for Each Category.

As can be seen in Figure VIII all the categories have different median durations but typically stay between 100 and 200 seconds, with the exception of the 'Pets & Animals' category being around 80 seconds. Looking at Figure VII the most notable trend is that the category 'Howto & DIY' has a much higher average video duration than other categories at around 600 seconds, which is well over two times more than the second highest category. Using this observation in conjunction with the previous observation of Figure VIII it can be concluded that since 'Howto & DIY' has a similar median to other categories but a much larger average it must have a similarly small portion of videos that are shorter and significantly longer than around 200 seconds. Which would explain a similar median but a different average video duration. When taking into context what these videos are it makes sense for this trend to arise as this category is an educational category and unlike music videos, is not constrained by industry standards, and instead is impacted by the particular video topic.

D. What did the best 2007 videos look like?

The final question about the 2007 Youtube data that the team asked concerned what were the most popular videos and what could be derived from categorical statistics on these videos to make a video popular. For the sake of this paper popularity was measured only by the number of views a video received. For the first part of analysis it was asked what the distribution of the categories was of the top 100 videos in the dataset. The distribution of the categories is depicted in the pie chart in Figure IX.

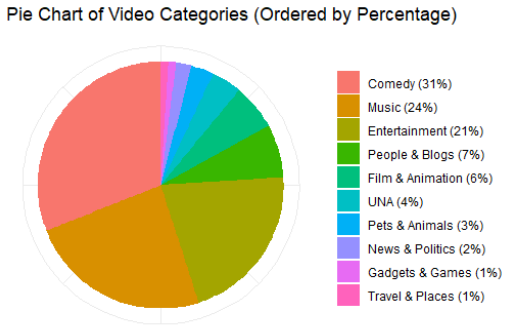

Figure IX: Pie Chart of the Categories for the Top 100 Videos.

As can be seen in Figure IX above Comedy videos comprised 31% of the top 100 videos, with Music comprising 24% and Entertainment at 21%. This closely resembles our overall distributions of views for each category seen back in Figure I. This allows us to conclude that the top 100 videos distribution closely matches the distribution for the overall view counts of each category. This is likely contributed to the cases previously noted where outlier viral videos tend to skew and data in a way that makes them quite noticeable. One significant difference that this graph shows is that 'UNA' only comprise 4% of the top 100 videos, but given Figure V and Figure VI it shows that 'UNA' videos typically get more views than other categories. This indicates that the 'UNA' category has significantly fewer outlier viral videos than the other categories.

For the final part of evaluation we wanted to look at the video lengths of the top 100 videos in each category to see if the average video length stayed the same and if the categories had similar distributions for video length for their most popular videos. This is depicted below in Figure X.

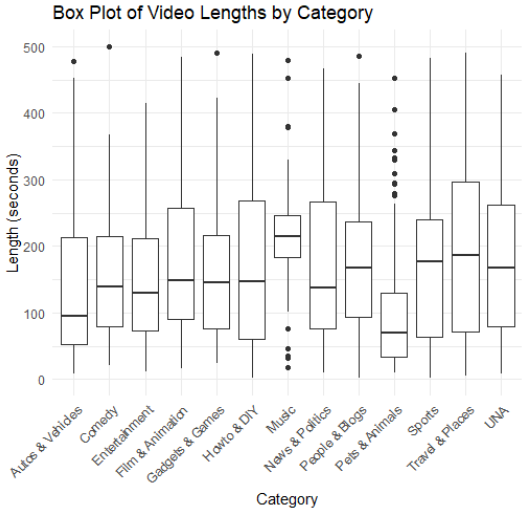

Figure X: Box Charts of the top 100 Videos for Each Category

From Figure X it can be seen by comparing across the various box plots, most categories have a similar average duration of around 150 seconds, with the exception of 'Autos & vehicles' and 'Pets & Animals' which are closer to 100 seconds. This leads to the conclusion that generally the most popular videos for each category average to around 150 seconds. The second notable observation is that the spread, or size of the box, varies widely between each of the categories. An exceptional category for consistency when it comes to this is 'Music' videos, but this is likely due to industry pressure for videos to be in a standard range, as the other categories are less impacted by outstanding entrenched industries. The least consistent categories include 'Howto & DIY' and 'Travel & Places'. Additionally, some categories have the average at an asymmetrical position in the box, likely due to a disproportionate number of videos being smaller or larger but being forced by various outliers that cause the average to not be symmetrical, and this is a point for further expiration in future projects. In conclusion, to make a popular video in 2007, it would be best to make a roughly 150 second Comedy or Entertainment video.

VI. CONTRIBUTIONS

During the process of implementation of R1, the group project was split into two parts. Jacob worked on the first half of the project, developing the command line interface to extract the data. Once the data was able to be extracted Gabriel and Tristan worked on implementing code that would represent the data as well as write about the significance of the results. Halfway through the project there was an idea for a GUI to seamlessly connect data collection and visualization however due to time limitations this idea was ultimately scrapped.

Implementer	Responsibilities
Tristan Bailey (30%)	1. Categorical Stats 2. R programming for data visualization concerning ½ of material. 3. Pyspark Methods for Acquiring some Categorical Statistics
Gabriel Mortensen (30%)	1. Menu rough draft 2. GUI prototype 3. R programming for data visualization concerning ½ of material 4. SQL Option to CLI for additional queries
Ean Jacob Gayban (40%)	1. CLI functionality 2. Pyspark integration 3. Bin Generation 4. Github Installation 5. Coded final draft for most Spark functionalities

Table IX.A: Table consisting of the responsibilities of each team member of R2-Team-1

VII. Related Work

Due to the abundance of data available on YouTube, researchers commonly analyze video metrics for research purposes. One related work that explores the use of YouTube data and analysis in Hadoop is the "YouTube Data Analysis Using Hadoop" by Charu Khosla [6]. In this work, Khosla mined data from YouTube to showcase how it can be utilized to make real-time and targeted decisions. Although our project utilized a dataset from 2007, Khosla's work could be applied for small-scale studies on specific channels and genres. The use of Hadoop and HDFS in Khosla's work is similar to our project's use of Apache Spark. While Khosla also explored real-time analysis, which we could not perform due to the age of our dataset, our tool allows for formatted data to be loaded and analyzed. Therefore, Khosla's work could be a possible next step for our project, as it delves into more in-depth analysis of the data.

VIII. Conclusion and Takeaways

The team demonstrated a way to query a large set of YouTube data utilizing PySpark then utilized R for data visualization. Although this project was not a new experience for anyone it reminded the team about the importance and efficiency of PySpark as well as other big data tools. A few key takeaways from the project include the following:

Popularity Factor: Based on the analysis of the YouTube dataset, we can draw some key takeaways about the factors that make a video popular and how this could impact society. For instance, since music videos are the most popular category on YouTube, businesses looking to reach a wider audience may consider advertising in this category. Similarly, since comedy videos are also popular, advertisers could benefit from placing their ads on these videos to gain more attention. Additionally, understanding the types of videos that are popular could help businesses and organizations identify potential trends and opportunities. However, it's important to note that while popular videos may have a significant impact on society, it's essential to promote positive trends that benefit society rather than harmful ones. By analyzing video statistics, we can identify potentially dangerous trends and take proactive measures to prevent negative outcomes.

Acknowledgment

We would like to thank Dr. Lei Yang for his knowledge and expertise in teaching the CS631 Big Data Systems class.

References

[1] Yeung D. (2018). Social Media as a Catalyst for Policy Action and Social Change for Health and Well-Being: Viewpoint. Journal of medical Internet research, 20(3), e94.

[2] https://netsg.cs.sfu.ca/youtubedata/

[3] Robson. (2019). Dangerous detergent : dealing with the Tide Pod challenge. SAGE Publications: SAGE Business Cases Originals.

[4] "Power Law Distribution." *Power Law Distribution - an Overview | ScienceDirect Topics*, https://www.sciencedirect.com/topics/mathematics/power-law-distribution.

[5] Easley, David, and Jon Kleinberg. "Chapter 18.2 Power Laws - Cornell University." *Networks, Crowds, and Markets*, https://www.cs.cornell.edu/home/kleinber/networks-book/networks-book-ch18.pdf.

[6] Khosla, C. (2016). *YouTube Data Analysis Using Hadoop*. scholarworks.calstate.edu. Retrieved April 23, 2022, from https://scholarworks.calstate.edu/downloads/k3569434b